Exploratory Data Analysis Report

1. Data Preparation Summary

The original dataset consisted of **128,975 records and 24 columns**. Several columns were either irrelevant to the analysis or contained a high percentage of missing values. The following data preparation steps were performed to ensure data quality and consistency:

- **Column Removal:**
  The columns *currency*, *ship-country*, *promotion-ids*, and *Unnamed: 22* were removed as they did not contribute meaningful information to the analysis.
- **Column Renaming:**
  To improve clarity and readability, selected columns were renamed:
  - SKU → *Stock Keeping Unit*
  - ASIN → *Amazon Standard Identification Number*
  - Qty → *Quantity*
  - B2B → *Business to Business*
- **Duplicate Records Check:**
  A duplicate check confirmed that there were **no duplicate rows** in the dataset.
- **Handling Missing Values:**
  Significant null values were observed in columns such as *Courier Status* (6,872), *Amount* (7,795), and *fulfilled-by* (89,698).
  To maintain analytical accuracy, **all rows containing null values were removed**, reducing the dataset to **32,395 records**.
- **Data Type Conversion:**
  The *Date* column was converted from object type to datetime format (%m-%d-%y) to enable time-series analysis.

After these steps, the dataset was clean, structured, and suitable for further analysis.

---

2. Sales Trends Analysis

Data visualizations were used to identify key sales patterns and trends:

- **Daily Sales Trend:**
  A line chart of *Total Sales Amount vs Date* showed fluctuating yet consistent sales activity over time. Peaks and dips suggest possible seasonal patterns or event-driven sales, which could be explored further.
- **Top Performing Product Categories:**
  Bar chart analysis revealed that **'kurta'** and **'Set'** categories significantly outperformed others in total sales amount, making them critical revenue drivers.

- **Top Shipping States:**
  Sales by shipping state indicated that **Maharashtra, Karnataka, and Tamil Nadu** were the top contributors to total revenue, providing valuable geographical insights for marketing and logistics planning.

---

3. Outlier Detection and Handling

Outliers were analyzed using **box plots** and the **Interquartile Range (IQR) method** for numerical columns:

- **Quantity Column:**
  The distribution showed a tight spread with a small number of outliers extending beyond the whiskers.
- **Amount Column:**
  A wider distribution with noticeable high-end outliers was observed.

IQR-Based Outlier Treatment:

- For the *Amount* column:
  - Q1 and Q3 were calculated
  - IQR was computed (Q3 − Q1)
  - Lower and upper bounds were defined using 1.5 × IQR
  - **1,146 outliers** were identified and removed
  - Dataset size after removal: **31,249 rows and 20 columns**
- For the *Quantity* column:
  - **128 outliers** were identified
  - These were noted but not explicitly removed unless overlapping with *Amount* outliers

Removing extreme outliers from the *Amount* column improved the robustness and reliability of subsequent statistical analyses.

---

4. Hypothesis Testing Summary

Hypothesis testing was conducted to compare **sales amounts between 'Set' and 'kurta' categories**.

Normality Test (Shapiro-Wilk):

- Both categories showed **p-values < 0.05**
- Conclusion: Sales amounts for both categories are **not normally distributed**
- This was supported by histogram and Q-Q plot visualizations showing skewness

Variance Homogeneity Test (Levene's Test):

- p-value < 0.05
- Conclusion: Variances between the two categories are **significantly different**

Independent Samples t-Test:

- T-statistic: **117.145**
- P-value: **0.000e+00**
- Result: A **highly significant difference** exists between the mean sales amounts of ‘Set’ and ‘kurta’

Although normality and equal variance assumptions were violated, the extremely small p-value indicates a strong and statistically meaningful difference. A Welch’s t-test would be more appropriate; however, the conclusion remains unchanged.

---

5. Final Conclusion

This comprehensive analysis of the **Amazon Sale Report dataset** provided meaningful insights into sales performance, product demand, and regional trends.

Key Findings:

- **Sales Performance:**
  ‘Kurta’ and ‘Set’ categories emerged as top revenue generators.
- **Geographical Insights:**
  Maharashtra, Karnataka, and Tamil Nadu were identified as the highest contributing states.
- **Data Quality Improvement:**
  Outliers in sales amount were effectively removed using the IQR method.
- **Statistical Evidence:**
  Hypothesis testing confirmed a significant difference in average sales amounts between ‘Set’ and ‘kurta’ categories.

Business Impact:

These insights can support:

- Inventory optimization for high-performing categories
- Targeted marketing strategies in top-performing states
- Deeper analysis of customer purchasing behaviour